PATENT ABSTRACTS OF JAPAN

(11)Publication number : 02-028769
(43)Date of publication of application : 30.01.1990

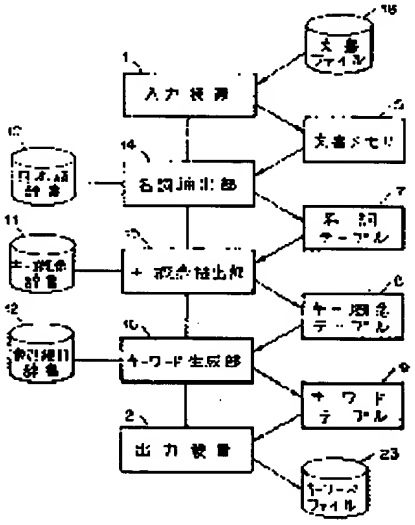
(51)Int.Cl. G06F 15/40

(21)Application number : 63-179802
(22)Date of filing : 18.07.1988

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>
(72)Inventor : NAGATA MASAAKI
KIMOTO HARUO

(54) AUTOMATIC KEY WORD GENERATING DEVICE

(57)Abstract:
PURPOSE: To generate a key word as a word to express the theme of a sentence by generating automatically the key word by using an indexing rule dictionary from a key conception extracted by a key conception extracting part.
CONSTITUTION: An inputting device 1 reads a document file 18 into a document memory 5, and a noun extracting part 14 extracts a noun by using a Japanese dictionary 10, and stores it in a noun table 7. Next, the key conception extracting part 15 collates a key conception dictionary 11, and enumerates all the key conception capable of being induced from the noun. The key conception is given a score from the intensity of relation to the induced noun, the position of appearance and the frequency of the appearance. The key conception whose score exceeds a threshold determined beforehand is stored in a key conception table 8. A key word generating part 16 collates the indexing rule dictionary 12, and enumerates all the key words stored in the table 8, and if they are constituted of a single key conception or if they are constituted of all the key conceptions, it outputs them.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

⑲日本国特許庁(JP)　⑪特許出願公開

⑫公開特許公報(A)　平2−28769

⑤Int.Cl.[5]	識別記号	庁内整理番号	㊸公開　平成2年(1990)1月30日
G 06 F 15/40	500 T	7313−5B	

審査請求　未請求　請求項の数 1　(全7頁)

㊹発明の名称　キーワード自動生成装置

㉑特　願　昭63−179802
㉒出　願　昭63(1988)7月18日

㉒発明者　永田　昌明　東京都千代田区内幸町1丁目1番6号　日本電信電話株式会社内
㉒発明者　木本　晴夫　東京都千代田区内幸町1丁目1番6号　日本電信電話株式会社内
㉑出願人　日本電信電話株式会社　東京都千代田区内幸町1丁目1番6号
㉔代理人　弁理士　草野　卓

明　細　書

1. 発明の名称

キーワード自動生成装置

2. 特許請求の範囲

(1) キーワードが表現する概念とこの構成要素となる基本概念(これをキー概念と呼ぶ)との関係を記憶する索引規則辞書と、

キー概念を想起させる能力を持つ単語集合を記憶するキー概念辞書と、

文章中から名詞を抽出する名詞抽出部と、

この名詞抽出部により抽出された単語に対応するキー概念を上記キー概念辞書を用いて抽出するキー概念抽出部と、

このキー概念抽出部により抽出されたキー概念から上記索引規則辞書を用いてキーワードを生成するキーワード生成部とを備えたキーワード自動生成装置。

3. 発明の詳細な説明

「産業上の利用分野」

この発明は文書データベース作成のために、データベースに蓄積される文書に対して、文章の内容を適切に表現するキーワードを文章中から自動的に生成するキーワード自動生成装置に関するものである。

「従来の技術」

新聞記事、特許明細書、技術論文などの大量の文書を含むデータベースを作成する場合、データベースへの入力の際に各文書に対して検索用のキーワードを付与しなければならない。従来、この目的のために次のような方法が用いられていた。

〔a〕単語の頻度による方法

〔b〕不要語辞書を用いる方法

〔c〕キーワード辞書を用いる方法

c1)完全一致法

c2)部分一致法

しかし、これらの従来の方法にはそれぞれ次のような欠点がある。

〔a〕の方法では、対象文章中の単語の出現頻度を調べ、出現頻度が中程度の語が文章の特徴を最もよく表していると考え、これをキーワードとす

る。この方法では高頻度語は一般的な用語とみなして除去し、低頻度語は特殊な用語とみなして除去する。このためキーワードとして抽出された語には統計的な文書識別能力が保証されるという利点がある。しかし文章を統語的あるいは意味的には解析しないので、抽出されたキーワードは必ずしも文章の主題を表現する語ではない。従って人手によりキーワードを抽出する場合に比べると、キーワードとなり得ない語を抽出することによる適合率の低下、キーワードとなり得る語を除去することによる再現率の低下が問題となる。

〔b〕の方法では、形容詞、形容動詞、副詞やキーワードになり得ない動詞、名詞などを収集した不要語辞書を予め用意し、対象文章中の単語とこの不要語辞書とを照合して、一致しなかった語をすべてキーワードとする。このため文章中に現れたキーワードとなり得る語を除去することにより再現率が低下することはない。しかし人手によりキーワードを抽出する場合に比べると、文章の主題とは余り関係のない語が大量にキーワードとして抽出されてしまうために適合率が大きく低下するという問題がある。またキーワードとして用いられる用語が統制されていないので、表記の違いや同義語による再現率の低下も問題となる。

〔c1〕の方法では、キーワードになり得る語を収集したキーワード辞書を予め用意し、対象文章中の単語とこのキーワード辞書を照合して、一致した語をキーワードとする。このためキーワードになり得ない語が抽出されることにより適合率が低下することはない。またこの方法ではキーワードとして用いられる用語が統制できるという利点がある。しかしキーワード辞書中の語が文章中に出現すれば自動的に抽出されるので、人手によりキーワードを抽出する場合に比べると、文章の主題に余り関係のない語が抽出されることにより適合率が低下するという問題がある。さらにキーワード辞書中の語と文章中の語が文字列として完全に一致しないと抽出されないので、対象とする文章の主題を表現するのに適切なキーワードがキーワード辞書中に存在してもキーワードとして抽出されないという問題がある。

〔c2〕の方法では、キーワード辞書を用いる点は〔c1〕と同様であるが、対象文章中の単語とキーワード辞書を照合する際に、完全に一致する語の他に部分的に一致する語もキーワード候補として抽出し、その中で一致度の高い語をキーワードとする。このため〔c1〕と比べると、キーワード辞書に収録されている語が変形した形（派生、省略、複合語化、分割など）で対象文章中に出現する場合でもキーワードを生成できるという利点がある。しかし文字列上の一致度は必ずしも意味的な類似度に対応していないので、文章の主題とは全く無関係なキーワードを生成してしまうことがあるという問題点がある。

これらをまとめれば、〔a〕〔b〕〔c〕の方法及びこれらを組み合わせた方法では、文章全体の意味的な解釈を行わないので、主題を表現するのに適切なキーワードを生成できない、主題と余り関係のないキーワードを生成してしまうという問題がある。特に〔a〕〔b〕〔c〕の方法では、文章中に出現しない語をキーワードとして生成することができない。また〔c1〕の方法では、文章中に出現しない語もキーワードとして生成できるが意味的な根拠が希薄である。

この発明の目的は、従来の方法では、文章の主題を表現するのに適切でない語が文章中からキーワードとして抽出されるという問題点や文章の主題を表現するのに適切な語が文章中に出現しなければ、キーワード辞書中に適切な語がある場合でも、キーワードとして生成されることはないという問題点を解決したキーワード自動生成装置を提供することにある。

「課題を解決するための手段」

この発明は、キーワード辞書中のキーワードは一つの概念を表す幾つかの用語の中から一つの用語だけを代表として選んで収録したものであるという性質、及び実際の文章中におけるこの概念の表層的な表現形態は、キーワード自身による場合、キーワードの同義語や厳密な意味では同義語ではないがキーワードが表す概念と同じ概念を喚起す

る能力を持つ語（広義の同義語）による場合、キーワードの表す概念を直接的に指示しないが、この概念を強く連想させる能力を持つ語（広義の関連語）による場合、に分類できるという性質を利用して、文章中の表層語とキーワードが表す概念（またはこれを構成する基本概念）の関係をキー概念辞書中に記述したこと、

文章の主題を表現する語としてあるキーワードが選ばれる場合、そのキーワードが表す概念（またはこれを構成する基本概念）が上述のいずれかの形で文章中に出現するという性質を利用して、キー概念辞書を用いて、文章中から抽出した名詞を調べることにより、文章中に出現する重要な概念（キー概念）を抽出すること、

キーワードが表現する概念が複合概念である場合、これを基本概念（キー概念）の組み合わせとして表すことができるという性質を利用して、キーワードとキー概念の関係を索引規則辞書中に記述したこと、

複合概念を表現するキーワードが文章の主題を

表現するのに適切であるときには、複合概念を構成する各基本概念が文章中に出現するという性質を利用して、索引規則辞書を用いて、文章中から抽出した概念の組み合わせを調べることにより、文章全体の主題を表現するキーワードを生成すること、

を最も主要な特徴とする。

従来の技術とは、広義の同義語及び広義の関連語からなるキー概念辞書を用いているので、文章中にキーワード辞書と完全に一致する語が出現しない場合でも、概念を抽出して適切なキーワードを生成できること、

キーワードが表現する概念が、文または文章全体の内容の解析を必要とするような複合概念である場合でも、個々の基本概念を抽出しその組み合わせを調べることにより適切なキーワードを生成できること、

が異なる。

「実施例」

第１図はこの発明の一実施例のシステム構成図

である。同図において１は磁気記憶装置に文字コードで記録されている文書データを読み込む入力装置、２は生成されたキーワードを磁気記憶装置に出力する出力装置、３はキーワード生成のプログラムを実行するプロセッサ（ＣＰＵ）、４はキーワード生成のプログラムを格納するプログラムメモリ、５は入力装置１により読み込まれた文書データを格納する文書メモリ、６はキーワード生成のプログラムを実行する際に使用する作業メモリ、７は文章から抽出した名詞を格納する名詞テーブル、８は名詞から抽出したキー概念を格納するキー概念テーブル、９はキー概念から生成したキーワードを格納するキーワードテーブル、１０は文章から名詞を抽出する際に必要な語彙情報と文法情報を格納した日本語辞書、１１は名詞とキー概念の関係を格納したキー概念辞書、１２はキー概念とキーワードの関係を記述した索引規則辞書である。

第２図はこの発明の一実施例の機能ブロック図である。入力装置１は処理対象となる文書ファイル１８を文書メモリ５に読み込む。次に名詞抽出部１４は日本語辞書１０を用いて対象とする文章から名詞を抽出し、名詞テーブル７に格納する。次にキー概念抽出部１５はキー概念辞書１１を照合し、名詞テーブル７に格納されている名詞から同義語または関連語の関係により導出可能なキー概念をすべて列挙する。列挙されたキー概念は次の３つの基準を用いて得点付が行われる。

1) キー概念を導出した名詞とキー概念の関連の強さ（同義語または関連語）

2) キー概念を導出した名詞の入力文章中の出現位置

3) キー概念を導出した名詞の入力文章中の出現頻度

異なる名詞から同じキー概念が導出される場合には、これらの得点を合計する。こうして各キー概念に対して得点が与えられ、この得点が予め決めたしきい値を越えたキー概念を入力文章から抽出されたキー概念としてキー概念テーブル８に格納する。次にキーワード生成部１６は索引規則辞書

１２を照合し、キー概念テーブル８に格納されているキー概念を構成要素として持つキーワードをすべて列挙する。列挙されたキーワードについて次の条件が満たされたとき、そのキーワードを入力文章に対するキーワードとしてキーワードテーブル９に格納する。

1) キーワードが単一のキー概念から構成されている

2) キーワードが複数のキー概念から構成され、構成要素となる全てのキー概念がキー概念テーブル８中に格納されている

最後に出力装置２はキーワードテーブル９に格納されているキーワードを外部記憶装置上のキーワードファイル２３に格納する。

第３図はキー概念辞書及び索引規則辞書の内容の一例である。キー概念は通常の名詞と区別するために／／で囲んである。第３図ａはキー概念／アメリカ合衆国／の同義語として「アメリカ」、「米国」、「合衆国」などの名詞が記憶され、関連語として「ワシントン」、「レーガン」などの

名詞が記録されていることなどを示す。第３図ｂはキーワード「米ソ関係」は、３つのキー概念／アメリカ合衆国／、／ソ連／、／関係／から構成されることを示す。

第４図はこの発明の一動作例である。入力装置１により文書メモリ５に読み込まれた入力文章27は名詞抽出部１４により名詞が抽出される。２８は名詞テーブル７の一部である。この例では冒頭の一文「ソ連のゴルバチョフ書記長は三十一日、モスクワで開かれたマシェル・モザンビーク大統領歓迎宴で演説し、・・・」という部分から、「ソ連」、「ゴルバチョフ」、「書記長」などの名詞が抽出されることを示す。次に名詞テーブル７の名詞からキー概念抽出部１５によりキー概念が抽出される。２９はキー概念テーブル８の内容である。この例では／ソ連／、／関係／、／アメリカ合衆国／などのキー概念が、第３図ａに示したようなキー概念辞書１１を用いて抽出されることを示す。

さらにキー概念テーブル８のキー概念からキー

ワード生成部１６によりキーワードが生成される。３０はキーワードテーブル９の内容である。この例では「米ソ関係」などのキーワードが、第３図ｂに示したような索引規則辞書１２を用いて生成されることを示す。比較のためにこの文章に対して人手により付けられたキーワードを３１に示す。ここで左端に「ｈ」を付けた語は自動生成されたキーワードである。

このような構成及び動作となっているから、文章中に現れるキーワードの同義語や関連語からキーワードが表す概念あるいはそれを構成する基本概念を文章中からキー概念として抽出し、キー概念の組み合わせを調べることにより文章全体の主題を表すキーワードを生成することができる。その効果としては従来の技術に比べて、文章中に出現しない語をキーワードとして生成することができ、また文章中に現れた概念の抽象化や組み合わせにより生ずる概念を表すキーワードを生成することができるという改善があった。

「発明の効果」

以上説明したように、キーワードが表す概念あるいはそれを構成する基本概念を対象とする文章中に現れるキーワードの同義語や関連語からキー概念として抽出し、キー概念の組み合わせを調べることにより文章全体の主題を表すキーワードを生成するのであるから、文章中に現れた概念を表すキーワードと同形の語が文章中に出現しない場合でも、キー概念辞書を用いて文章中の表層語から概念を抽出することによりキーワードを生成することができ、また文章中に現れた概念の抽象化や組み合わせを表すキーワードが必要な場合には、索引規則辞書を用いてキー概念の組み合わせを調べることにより生成することができるという利点がある。

4. 図面の簡単な説明

第１図はこの発明の一実施例のシステム構成図、第２図はこの発明の一実施例の機能ブロック図、第３図はこの発明で用いられる辞書内容の一例を示し、第３図ａはキー概念辞書の一部を示す図、第３図ｂは索引規則辞書の一部を示す図、第４図

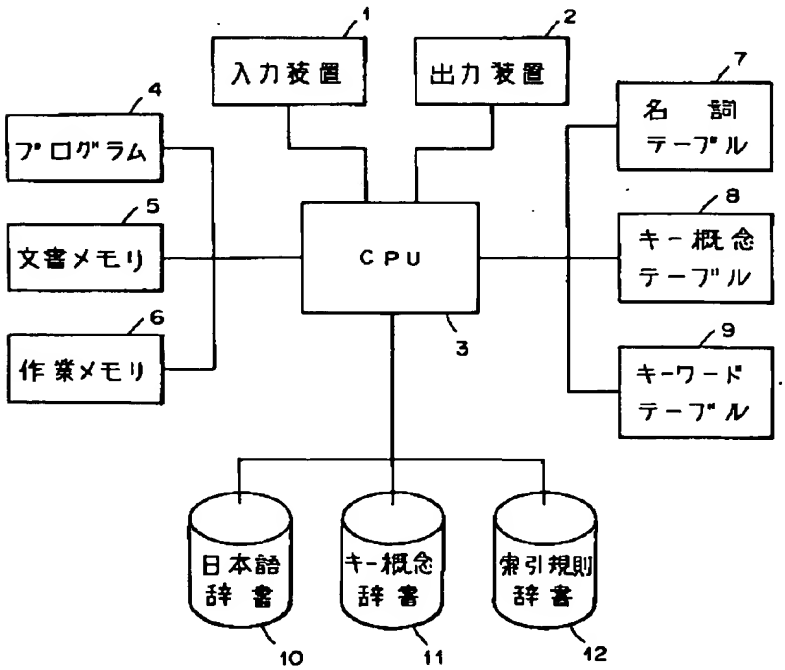
はこの発明の一動作例を示す図である。

特許出願人　　日本電信電話株式会社

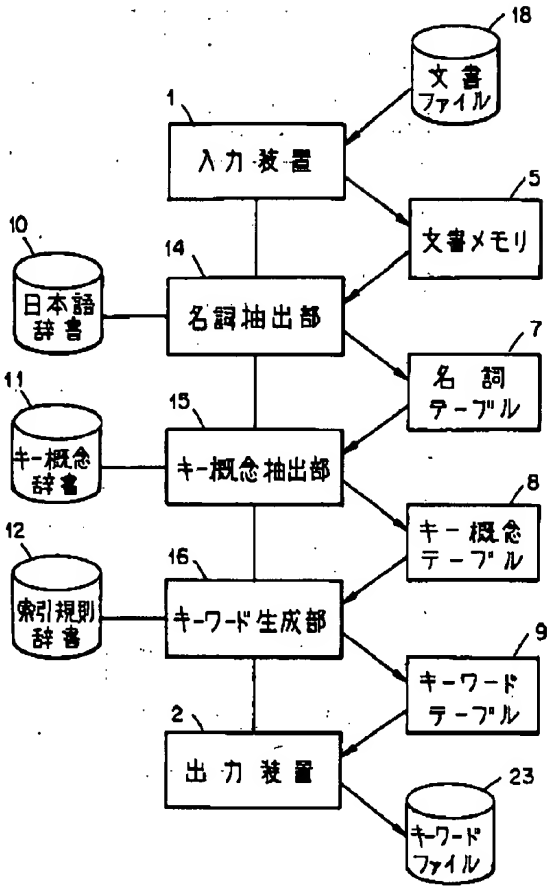
代　理　人　　草　野　　卓

(15)

第 1 図



第 2 図



第 3 図

(a) キー概念辞書(一部)

キー概念	種別	名　　詞
/アメリカ合衆国/	同義語	アメリカ，米国，合衆国，---
/アメリカ合衆国/	関連語	ワシントン，レーガン，---
/ソ連/	同義語	ソ連，ソビエト，---
/ソ連/	関連語	モスクワ，ゴルバチョフ，---
/関係/	同義語	関係，友好，外交，貿易，会談，対話，合意，条約，---

(b) 索引規則辞書(一部)

キーワード	キー概念
米ソ関係	/アメリカ合衆国/，/ソ連/，/関係/
自動車事故	/自動車/，/事故/
人事	/組織名/，/人名/，/役職名/，/人事用語/

特開平 2-28769(6)

第 4 図

27 原 文

核実験停止再度呼びかけ　ソ連書記長

【モスクワ三十一日大熊特派員】ソ連のゴルバチョフ書記長は三十一日、モスクワで開かれたマシェル・モザンビーク大統領歓迎宴で演説し、核実験停止に関するソ連提案に米国が前向きの姿勢を示すよう、再度呼びかけた。「ソ連は核実験停止について直ちに合意することを米国に提案した。今、自分善意を行動で示す現実的なチャンスがある。ワシントンからの責任あるアプローチを待っている」と述べたもの。

28 名詞（一部）

ソ連
ゴルバチョフ
書記長
モスクワ
マシェル

モザンビーク
大統領
歓迎宴
核実験
停止

29 キー概念

地名
ソ連
人名
人の属性
モザンビーク
大統領
関係
アメリカ合衆国

30 自動生成した キーワード

h ソ連
モザンビーク

大統領
h 米ソ関係

h アメリカ合衆国

31 人間が付けた キーワード

h ソ連
核実験
h アメリカ合衆国
h 米ソ関係
ゴルバチョフ，ミハイル